DOI:10.1145/3605160

Best practices for open source ecosystems researchers.

BY AMANDA CASARI, JULIA FERRAIOLI, AND JUNIPER LOVATO

Beyond the Repository

OPEN SOURCE IS much more than a repository—it is a rich, multilevel ecosystem of human contributors who collaborate and cooperate in many capacities to accomplish a shared creative endeavor. Consequently, when studying open source ecosystems, numerous interacting parts must be considered to understand the dynamics of the whole. Research on open source ecosystems is ultimately research about a sociotechnical ecosystem. Researchers should take care to retain the *socio* element in research and understand how both their methods and results may impact entire open source ecosystems.

This article describes best practices for open source ecosystems research through multiple overarching best practices. It offers practical guidelines for conducting rigorous, ethical, respectful research that maintains the integrity of the open source ecosystem under consideration.

Open source projects and ecosystems have evolved into a critical sociotechnical system underpinning much of modern society.[13] As a significant component of STEM (science, technology, engineering, and mathematics) fields, open source itself is studied under many disciplines, including science, economics, data ethics, computer science, psychology, sociology, and more.

Open source ecosystems benefit from comprehensive and scientifically sound examination by:

- Setting expectations for mutually beneficial best practices for critical vulnerability disclosure and remediation.[26]
- Analyzing ecosystem-level dynamics to identify localized factors impacting project life cycle development.[28]
- Identifying fine-grained development practices that have population-level effects on historically minoritized populations.

Quickly escalating situations can develop when open source projects and open data available about open source are viewed as "free" opportunities for research. When researchers fail to consider the human element of open source, it harms open source ecosystems by:

- Increasing demands on often-overwhelmed volunteer groups.[25]
- Impacting costly infrastructure systems not designed to support research use cases.[7]
- Treating vital open source systems as test beds for scholarly research into known problems without the consent of the community or contributing back to correct these problems.[6]

Much of the existing research about open source elects to study software repositories instead of ecosystems.[16] An open source repository most often refers to the artifacts recorded in a version control system and occasionally includes interactions around the repository itself. An open source ecosystem refers to a collection of repositories, the community, their interactions, incentives, behavioral norms, and culture. The decentralized nature of open source makes holistic analysis of the ecosystem an arduous task, with communities and identities intersecting in organic and evolving ways.

Despite these complexities, the increased scrutiny on software security and supply chains makes taking an

IMAGE BY LOGIN/SHUTTERSTOCK

ecosystem-based approach of utmost importance when performing research about open source, as illustrated in the accompanying figure. This article provides guidelines and best practices for research using data collected from open source ecosystems, encouraging research teams to work with communities in respectful ways.

1. Always treat open source ecosystems as systems "in production."

Open source touches mission-critical systems across the world and has an increasing impact on global populations.[19] Some of these systems cannot be patched outside of prescheduled maintenance windows, if at all. They may be what powers a person's insulin pump,[3] models water quality,[21] or determines eligibility for benefits of underserved people.[2]

The infrastructure and communities that make up open source ecosystems are not experimental test environments.[6] Running behavioral or technical experiments in open source ecosystems may impact the world's infrastructure in *unknown and immeasurable* ways. In some cases, this impact may cause real and lasting harm to both participants in the ecosystem and end users without any agency in the ecosystem. As framed in one public guideline: "Do not impact other users with your testing; this includes testing vulnerabilities in repositories or organizations you do not own."[15]

Additionally, as open source affects the lives of people without them knowing about it, getting informed consent for experiments (even with the best of intentions) is impossible. Open source ecosystems may appeal to researchers as prime candidates for study by virtue of how they embrace transparency in both process and outcomes. Still, researchers should consider these ecosystems to be perpetually "in production" and exercise extreme caution when designing experiments.

BEST PRACTICE: *If experiments have the potential to impact open source ecosystems or populations, ensure they are self-contained, have no side effects, and are reversible if needed.*

2. Assume the economic incentives and availability of the people who keep the lights on are not evenly distributed.

The composition of the open source ecosystem has shifted over the decades from primarily unpaid contributors to a mixture of unpaid and paid contribu-

Four overlapping areas of consideration.

When researching open source ecosystems, be cognizant of...

Data ethics

- Get consent from and consult with communities being studied if gathering data about people within the community
- Find the balance between privacy, ethics, and transparency when processing and sharing data
- Use data from opt-in sources instead of inferential methods

Research best practices

- Use community best practices and improve existing ecosystem programs
- Take into account social factors when scoping research questions
- Be clear and specific about observations, sampling methods, and data sources

Respect and equity

- Consider the impact of your research on populations within an ecosystem or sub-ecosystem
- Take into account the uneven distribution of incentives, economic and otherwise

Ecosystem integrity

- Treat open source ecosystems as perpetually "in production"
- Adhere to relevant licenses and assume the highest level of creator ownership over information and data

tors. Because of the disparity of direct financial support for contributors and maintainers alike, participation in open source varies greatly depending on economic factors.[5] Moreover, these economic incentives (and disincentives) may be absent or inconsistently observable from available data.[24]

Economic incentives may also not be directly observable at all. Given the propensity for companies to use open source participation for screening potential candidates, and similarly to use their own participation to recruit potential candidates,[18] contributors may view their participation as investments in their own financial and reputational interests. This also skews the work being done in open source toward work that can be measured or otherwise quantified; work such as design or accessibility testing often does not appear in available data and is therefore economically and reputationally disincentivized.[31]

BEST PRACTICE: *Include and document factors that preclude or disincentivize participation in open source ecosystems in any analysis of their populations.*[10]

These factors may be reflected in the data, or lack thereof, representing different types of (potentially overlapping) biases:

- *Compensation bias.* Participation without compensation is often not a viable option for many who have relevant skills or might be anomalous for their area of expertise.
- *Access bias.* Equipment and/or a reliable Internet connection may not be accessible outside the context of a person's employment.
- *Geographical bias.* Potential contributors may encounter barriers to participation, such as sanctions or discrepancies in working hours, by virtue of geographic location.
- *Unallocated time bias.* Individuals who have familial commitments, are disabled, or are experiencing economic hardship have less available time to dedicate to participation.

When analyzing the economics and sustainability of open source ecosystems, researchers must account for how the data selected may not paint a complete picture of the work being performed. Furthermore, they should carefully and explicitly examine how their conclusions may exacerbate inequalities in the open source ecosystem by placing disproportionate value on types of work simply because of convenient data.

3. Examine all information online in a way that honors attached licenses or assumes the highest level of creator ownership—for software, for data, for content, for all of it.

Published "publicly" does not automatically mean that information is available for reuse—whether in code, literature, or research. When researching open source ecosystems, ensure usage of the associated data follows the licensing and policies attached to the project or system to which it belongs.

Open source ecosystem data usage may be as straightforward as identifying the license attached to a repository.[23] With the increased use of platforms and third-party applications for open source community and infrastructure management, it is not safe to assume the source license attached to a project applies across all resources used by a community or for organizing open source work. Research each platform's terms and conditions and the specific community guidelines for each organization in these platforms.

For example, Wikimedia Commons specifically states its guidelines regarding photographs of identifiable people: "When dealing with photographs of people, we are required to consider the legal rights of the subject and the ethics of publishing the photo in addition to the concerns of the photographer and owner of the image."[8]

Open source communities also frequently build and host their own infrastructures using free or open source software, which have a combination of licenses and permissions about the software, data stored in, and community guidelines. This must all be untangled to discover the highest restrictive policy.

BEST PRACTICE: *If there is no guidance or document detailing how and under what conditions third parties may use the information, consider it unavailable for research purposes and exclude it from the data.*

4. Be clear and specific about your observations, sampling methods, and documentation of data sources.

While industry often uses the term *open source* to refer to the collective movement and development practice(s) of creating software adhering to the open source definition (OSD),[22] the overall open source ecosystem contains many sub-ecosystems that interact (or not) in complex ways. Given the ambiguity in language and potential for misinterpretation, research questions should identify the relevant population and sub-ecosystem with as much specificity as possi-

ble. Avoid generalizations about open source that may not apply to or be testable across sub-ecosystems.

Practices, communication mechanisms, and tooling in open source evolve rapidly and vary widely. This can lead to a greater-than-average heterogeneity of available data.

BEST PRACTICE: *When performing and publishing research based on quantitative data, document the parameters of the data collection process, including applicable time span, available data, and relevant population(s). Whenever possible, include or reference documentation*[14] *for the data collected or used.*

With few datasets available about open source, it is tempting to base new studies on preexisting or "convenient data" that may not be transferable across time frames, populations, or technologies. Be aware existing datasets may not be representative of the sub-ecosystem under consideration.[27] Use care when applying conclusions from one sub-ecosystem to another and ensure methodologies include testing and reestablishing of baselines.

Similarly, lean toward results that can be replicated per study rather than relying on conclusions drawn from prior studies. This minimizes the potential for recycling errors in methodology or amplifying bias in the open source ecosystem. Ensure data, code, and methodology are included in the publication of any interpretations or conclusions.

5. Use community best practices and improve exisitng ecosystem programs, even when this is not scientifically "novel."

There is no "silent" or "detached observer" role in open source. The impact researchers or analysts are able to have in a community, and the technosocial reputations they are building, will directly correlate to their consideration of how their work affects other people and their respective workloads. Each open source ecosystem has its own written and unwritten rules for submitting feedback, suggesting improvements, and identifying responsible parties when a technical emergency occurs. Researchers cannot ignore these norms, even when they are contrary to their scientific communities' pathways to promotion and achievement.

When analyzing the economics and sustainability of open source ecosystems, researchers must account for how the data selected may not paint a complete picture of the work being performed.

Scientific and industry research does not always reward the same kinds of findings useful for open source community members to improve their work and knowledge about their own spaces.

BEST PRACTICE: *Report any notable findings made during research using established community channels and practices, rather than discard problems that may not have novel scientific merit.*

Being part of internal feedback loops to improve open source with applied science not only contributes to the sociotechnical system researchers are a part of, but also increases technosocial trust between the researchers and their communities.

Privacy and security researchers have their own community guidelines and practices, which may not always carry over into general open source ecosystems. When research bleeds into new ecosystems, it is especially important to seek out communication norms and known issues. Most ecosystems have established programs with governing bodies to report vulnerability findings before open publication.[11] This applies to any publication, whether a blog or scientific paper, and prevents communities from feeling targeted by researchers.

6. Seek ground truth data from opt-in sources rather than from inferential methods.

Open source lacks basic social science baseline population data against which to compare research. This can lead to a lack of accountability when presenting survey results or when choosing unsupervised methodologies that allow for conclusions to be drawn with no comparative ground truth to validate the findings.

BEST PRACTICE: *Avoid algorithmic methods that are scientifically unsound and may further alienate or erase subpopulations when examining social structures in open source.*[1]

Researchers must be conscientious about steps taken during data selection, cleaning, and feature identification, specifically about categorization and easy reductionism of identities

to single categories. For example, the open source contributions people make as part of paid work may not be the same contributions and spaces they work in when not representing their employers. People do not readily sit in a single dimensional cluster.

The context under which a person participates in a specific ecosystem should be protected and not expected to carry across systems. Identity may be temporal—repeatedly asking about an individual's identity is important because these self-chosen words may change over time and in different areas of contribution. Allowing people to choose their own demographic and identity markers, given a specific context, lets individuals be represented most accurately as themselves.

Because many open source communities work with multiple information sources, it can be challenging to associate all work and data with specific individuals. Identity, however, can be specifically a factor of community, platform, or space. Individuals may be able to represent themselves in some communities without fear, but not in others. Because of this, each researcher needs to consult their own discipline's data ethics practices before performing anti-aliasing in data systems when subjects cannot self-identify in large-scale surveys.

7. Retain the socio-element when scoping research questions of a sociotechnical ecosystem.

When scoping a research question, do not reduce the problem into merely the technical elements and artifacts produced by a collaborative community and the human impacts on the ecosystem dynamics. Open source is a multidimensional ecosystem that involves an interplay between the interconnected network of technical and social systems. Using mixed-methods approaches and multidisciplinary collaborations can help unveil these critical dimensions in the systems under examination.

If you are seeking to understand a system of open source software artifacts, it may be equally important to understand the context in which it was created, and much of that context may not be visible in the code or metadata in the system. Understanding who created the artifacts, by what means, and under what circumstances, is a crucial aspect of understanding the quality and structure of the resulting product. It is also important to consider social context in order to avoid missing key variables that impact the structure (or structures) and dynamics of the sociotechnical ecosystem.[12,20]

Each open source ecosystem has its own written and unwritten rules for submitting feedback, suggesting improvements, and identifying responsible parties when a technical emergency occurs.

BEST PRACTICE: *Collaborate with members of the open source community, partners, and social scientists who can investigate the contextual frameworks behind the outputs in order to fully capture any open source ecosystem and avoid dark data problems in research results.*[17]

8. Get consent from and consult with communities being studied if gathering data about people in the community.

When beginning data collection, involve members of the affected community in the research life cycle from the design process to implementation, by communicating any findings. Consulting with the community aids in understanding the open source ecosystem, recruiting participants, validating results, and understanding the potential impact of the results on the community. It is also important to consult with community partners to communicate the findings at the end of the research pipeline to avoid misrepresentation of the community.

Researchers collecting new information (original data) about open source communities should get direct informed consent from the individuals and groups implicated in the research study. Whenever possible, enroll members into the study directly through recruitment and in cooperation with the relevant research ethics committee(s) or IRB (institutional review board) instead of simply scraping data from an online platform. Even if this is allowed by the terms of service (ToS), researchers should be held to a higher standard. Data concerning open source ecosystems implicates human subjects, and even if considered public, information collected from online open source communities may still contain sensitive personally identifiable information (PII).[9]

If using secondary data about open source communities (data not collected directly but obtained from a secondary source), be sure to check that the original data collector obtained informed

consent to collect the original dataset and that third parties have permission to use the data. While a dataset's metadata may indicate conditions under which others may use it, verify that permission by contacting the original data collector directly whenever possible.

BEST PRACTICE: *Obtain informed consent for any original data gathered during the research life cycle; understand how secondary data was collected, when it was collected, by whom, and under what oversight.*

9. Find the balance between privacy, ethics, and transparency when processing and sharing data.

While data ethics and open data can be seen as being at odds, both play an important role in research on open source ecosystems. Ethical collection and processing of data ensures the data is both methodologically and socially sound, while open data ensures the research can be reproduced and interrogated by the open source and research community. When possible, consider the FAIR (findability, accessibility, interoperability, and reusability)[30] and CARE (collective benefit, authority to control, responsibility, and ethics) principles[4] in concert with one another to find a balance between open and ethical data practices.

BEST PRACTICE: *Clearly label data that potentially implicates human subjects in metadata (for example, datasheet, code book, data cards, clear README file) to ensure transparency.*

Data that might not be considered sensitive alone can become sensitive if combined with another dataset. Aggregation of data can de-identify a dataset or expose sensitive information about an individual or group.[29] If combining datasets, take special care not to expose sensitive personal information by combining aliases that were intended to represent separate personal identities.

Openness is not necessarily a good in itself. If it is possible the data collected or aggregated as part of research into an open source ecosystem may implicate a community in a harmful or unethical way, researchers have a duty to consult with the implicated community, the IRB, and data ethicists to find the best way to make the study reproducible without releasing raw data.

Conclusion

Open source ecosystems extend far beyond the repositories they produce. Research into these ecosystems must account for the complex nature of open source, including the consideration of the downstream impact of the research itself. Researchers should keep data ethics, research best practices, respect and equity, and ecosystem integrity at the forefront of their minds while scoping, planning, executing, and publishing their research.

While this article outlines a number of considerations that fall into these themes, the complexity, scale, and rapid growth of open source ecosystems necessitate an evolving approach to research and may result in the formation of additional recommendations. These additions or expansions will likely still fall in one or more of the overarching themes, and contextualizing them in this manner may be helpful for future research and researchers. Open source ecosystems and the practice of research into them will continue to benefit from thoughtful and conscientious methodologies that incorporate these best practices. C

References

1. Aguera y Arcas, B. et al. Physiognomy's new clothes. *Medium*. (2017); https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a.
2. Assistant Secretary for Public Affairs. 3.14 open-source software. Department of Health and Human Services, (2016); https://www.hhs.gov/open/2016-plan/open-source-software.html.
3. Burnside, M.J. et al. 286-OR The CREATE trial: Randomized clinical trial comparing open-source automated insulin delivery with sensor augmented pump therapy in type 1 diabetes. *Diabetes 71*, (2022), Supplement_1; https://bit.ly/3N95O7R.
4. Carroll, S.R. et al. The CARE principles for indigenous data governance. *Data Science J. 19*, 11 (2020), 43; https://datascience.codata.org/articles/10.5334/dsj-2020-043/.
5. Carter, H., and Groopman, J. Diversity, equity, and inclusion in open source. *Linux Foundation*. (2021); https://bit.ly/3PdXNRO.
6. Chin, M. How a university got itself banned from the Linux kernel. *The Verge*. (2021); https://bit.ly/3pbLniD.
7. Clark, M. Security researcher finds a way to run code on Apple, PayPal, and Microsoft's systems. *The Verge*. (2021); https://bit.ly/3PaovLa.
8. *Commons: Photographs of Identifiable People*. Wikimedia Commons, (2023).
9. Computer Security Resource Center. Personally identifiable information. *Glossary*. National Institute of Standards and Technology; https://csrc.nist.gov/glossary/term/personally_identifiable_information.
10. D'Ignazio, C., and Klein, L.F. Strong ideas series. *Data Feminism*. MIT Press, Cambridge, MA (2020).
11. Django. Django's security policies; https://docs.djangoproject.com/en/4.1/internals/security/.
12. Dunne, J.A. et al. The roles and impacts of human hunter-gatherers in North Pacific marine food webs. *Scientific Reports 6*, 21179, (2016); https://www.nature.com/articles/srep21179.
13. Eghbal, N. Roads and bridges: The unseen labor behind our digital infrastructure. *Ford Foundation*. (2016); https://bit.ly/42OsdwZ.
14. Gebru, T. et al. Datasheets for datasets. *Commun. ACM. 64*, 12 (Dec. 2021), 86–92; https://cacm.acm.org/magazines/2021/12/256932-datasheets-for-datasets/abstract.
15. GitHub Security. GitHub bug bounty (2022); https://bounty.github.com.
16. Gold, N.E., and Krinke, J. Ethics in the mining of software repositories. *Empirical Softw. Eng. 27*, 1 (2021); https://dl.acm.org/doi/10.1007/s10664-021-10057-7.
17. Hand, D.J. *Dark Data: Why What You Don't Know Matters*. Princeton University Press, Princeton, NJ (2020).
18. Lerner, J., and Tirole, J. The simple economics of open source. *Harvard Business School Working Paper SSRN Electronic J.* (2000); https://papers.ssrn.com/sol3/papers.cfm?abstract_id=224008.
19. Lifshitz-Assaf, H., and Nagle, F. The digital economy runs on open source. Here's how to protect it. *Harvard Business Review* (Sept. 2, 2021); https://bit.ly/3JkVBEj.
20. Nissenbaum, H. *Privacy in Context*. Stanford University Press, Redwood City, CA, USA (2009), 186–230.
21. Office of Research and Development, USEPA. EPANET. (2014); https://www.epa.gov/water-research/epanet.
22. Open Source Initiative. The open source definition. (2007); https://opensource.org/osd.
23. Open Source Initiative. (2023); https://opensource.org.
24. Riehle, D. The economic motivation of open-source software: stakeholder perspectives. *Computer 40*, 4 (2007), 25-32; https://dl.acm.org/doi/10.1109/MC.2007.147.
25. Ruby, S. Confluence mobile position paper. Apache Software Foundation. (2022); https://bit.ly/3XcmrUI.
26. Schoen, R. A walk through project zero metrics. *Google Project Zero Blog*. (2022); https://bit.ly/3PiyIoO.
27. Trujillo, M.Z. et al. The penumbra of open source: projects outside of centralized platforms are longer maintained, more academic and more collaborative. *EPJ Data Science 11*, 31, (2022); https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-022-00345-7.
28. Valiev, M. et al. Ecosystem-level determinants of sustained activity in open-source projects: a case study of the PyPI ecosystem. In *Proceedings of the 26th ACM Joint Meeting on European Softw. Eng. Conf. Sym. Foundations of Softw. Eng.* (2018), 644–655; https://dl.acm.org/doi/10.1145/3236024.3236062.
29. Wagner, I., and Boiten, E. Privacy risk assessment: from art to science, by metrics. In *Data Privacy Management, Cryptocurrencies and Blockchain Technologies*. Springer International Publishing, (2018).
30. Wilkinson, M.D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data 3*, 160018 (2016); https://www.nature.com/articles/sdata201618.
31. Young, J.-G. et al. Which contributions count? Analysis of attribution in open source. In *Proceedings of the IEEE/ACM 18th Intern. Conf. Mining Software Repositories*. (2021); https://ieeexplore.ieee.org/document/9463079.

Amanda Casari is a researcher and engineer in the Open Source Programs Office at Google, where she is co-leading research and engineering to better understand risk and resilience in open source ecosystems. In 2021, she was named an External Faculty member of the Vermont Complex Systems Center, University of Vermont, Burlington, VT, USA.

Julia Ferraioli is an independent open source strategist, researcher, and practitioner with a decade of experience in launching, managing, and optimizing open source projects at scale. Her community work includes co-leading Open Source Stories, a Seattle, WA-based, community-led effort with the goal of making the people of open source and their lived experiences more visible.

Juniper Lovato is an educator and researcher in the field of complex systems and data science. She is the Director of Partnerships and External Programs, Vermont Complex Systems Center and a Ph.D. candidate in Complex Systems & Data Science at the University of Vermont, Burlington, VT, USA.